STUDIES ON STRUCTURAL EQUATION MODELING

Final Synopsis of the Doctoral Research work

Submitted to

MANGALORE UNIVERSITY

For the award of the degree of

Doctor of Philosophy (Ph.D.) in Statistics

By

SATYANARAYANA

Under the guidance of

Prof. ISMAIL B.

Department of PG studies and Research in Statistics

Mangalore University

Mangalagangothri-574199

Karnataka, India

MARCH 2023

INTRODUCTION

In economics and health science there are situations that involve more than one dependent variable and each dependent variable simultaneously depends on the remaining dependent variables in addition to the regressors in the form of a system of simultaneous equations. The parameters of any single structural equations are estimated using the Instrumental variable method. An estimator based on instrumental variables includes Indirect least squares, 2 Stage Least Squares (2SLS), Least Variance Ratio (LVR) and k-class estimators.

In health science, it is obvious that the variables are naturally related. i.e., the variables not only have a direct effect but also have some indirect effects on the study variable. Therefore, it is needed to model the exact structure of the relationship that exists between the variables simultaneously rather than carrying out separate regression analysis. When the outcome variable is dichotomous, the logistic regression model is commonly used to model the outcome variable. The main limitation of this model is that it considers only the additive nature of the variables. Therefore, there is a need to use a technique that can handle the most complex nature of the variables simultaneously and more effectively. Structural equation modelling (SEM) is a very powerful statistical tool for dealing with the above issues.

Through SEM , it is possible to analyze, simultaneously, both the relations of dependence between the Latent Variables (i.e., Structural Model) and the links between the Latent Variables and their indicators, that is, between the corresponding manifest variables(i.e., Measurement Model).SEM seeks to determine the extent to which the postulated structure is actually consistent with the empirical data at hand. This is carried out by computing the implied covariance matrix by means of the specified model, and comparing it to the (actual) covariance matrix based on the empirical data. The SEM model is given by

$$\underset{(m\ x\ 1)}{\eta} = \underset{(m\ x\ m)\ (mx1)}{\mathrm{B}\,\eta} + \underset{(m\ x\ n)(n\ x\ 1)}{\Gamma\,\xi} + \underset{(m\ x\ 1)}{\zeta} \qquad (1)$$

Where B and Γ are two matrix of structural coefficients between the endogenous variables, and between the exogenous and endogenous variables, respectively. The *measurement model* for endogenous variables is:

$$\underset{(px1)}{Y} = \underset{(pxm)}{\wedge_y}\ \underset{(mx1)}{\eta} + \underset{(px1)}{\varepsilon} \qquad (2)$$

where Y, η , ε are three vectors of the observed endogenous variables, latent endogenous and errors, respectively. $\wedge_y$ is the matrix of the structural coefficients between the observed variables and the latent variables. The *measurement model* for exogenous variables is:

$$\underset{(qx1)}{X} = \underset{(qxn)}{\wedge_x}\ \underset{(nx1)}{\xi} + \underset{(qx1)}{\delta} \qquad (3)$$

where $X, \xi\ and\ \delta$ are three vectors of the observed exogenous, latent exogenous and errors, respectively. Also, $\zeta,\ \varepsilon\ and\ \delta$ are mutually uncorrelated with $E(\zeta) = 0, E(\varepsilon) = 0, E(\delta) = 0$, Cov $(\zeta) = \Phi$, Cov $(\varepsilon) = \theta_\varepsilon$ and Cov$(\delta) = \theta_\delta$; ζ is uncorrelated with ξ, ε is uncorrelated with η and δ is uncorrelated with ξ. The regression analysis, canonical correlation, confirmatory factor analysis and repeated measures analysis of variance are the special cases of SEM. Regression models connect dependent variables with independent variables in an additive manner. Path analysis can be used to analyze the models that are more complex than the multiple regression models. The path model evaluates the effects of a set of variables on a specified outcome via multiple causal pathways. Path model ignores the measurement error in the variables whereas SEM uses latent variables to account for measurement error.

Machine learning models are constructed based on experiences learned from the training data set and it is used to extract hidden information from large volume of data. The Logistic regression model, Support Vector Machine (SVM), K-Nearest Neighbours models (KNN) and Regression tree are the leading models. The nature of data set model parameter selection and validation of results plays vital role in the model performance.

The estimation of interaction effects has expanded more prominence as they are often used by to identify the heterogeneous treatment effects. Regression tree can be used to model interaction effects. The main advantages (Fritz Schittz; 2018) of using a regression tree to model interaction effects are: (i) Tree-based methods include interactions by construction, without requiring the researcher to have any preconception on this matter which incapacitates the common specification errors. (ii) Nonlinear interaction effects are more naturally accommodated by tree-based methods, as contrasting to multiplicative interactions. (iii) Regression tree plots visualize interactions between variables.

While dealing cross-section data and time series data, we cross the problem of heteroscedasticity and autocorrelation respectively in the regression model. If the error variance is not the same across the observations then such a situation is known as heteroscedasticity. In the presence of heteroscedastic errors, OLS estimators are not efficient and the usual test procedures lead to the improper conclusion. It may also lead to a wider confidence interval which increases the risk of type II error. In this situation, the Generalized Least Squares Estimator (GLSE) can be used which is not only unbiased but also efficient.

Autocorrelation in regression models affects both the estimation and test procedure, so it is important to detect this problem. The Durbin–Watson (D-W) test is one of the most widely used test for auto-correlation. The major drawback of this test is that the test fails when DW statistic falls in the inconclusive region. As a solution to this problem, many studies have been carried out to remove the indeterminate range and improve the power of the DW test. Two approaches namely approximated distribution approach and the numerical computation approach are commonly seen in the literature. The power improvement by the approximated distribution approach is relatively smaller compared to the numerical computation approach. The numerical computation approach is to directly calculate the critical values by numerically

inverting the characteristic function of the rejection rule. However, in this approach critical values depend on the distributional assumption and the design matrix.

LITERATURE REVIEW

Two Stage Square (2SLS) estimator was introduced more or less independently by Theil (1953a, 1953b, 1961), and Sargan (1958). In particular, Anderson and Rubin (1949, 1950) developed the limited information maximum likelihood (LIML) estimator for the parameters of a single structural equation. Anderson (2005) gives the history of 2SLS a revisionist twist by pointing out that Anderson and Rubin (1950) indirectly include the 2SLS estimator and its asymptotic distribution. Through Structural Equation Models (SEM), it is possible to analyze simultaneously, both, the relations of dependence between the latent variables and the link between the latent variables and their indicators (Joreskog:1970). Joreskog and Sorbom (1993) introduced 2 SLS/ instrumental variable (IV) methods to estimate the model parameters of latent variable (structural) model.

The review paper of Ibomoiye et al (2019) highlights the prediction performance of various decision tree algorithms and they also carried out an in-depth review of relevant research that attempted to improve the performance of the algorithms and various methods used. Mitchell's (1997) study shows that redundant variables influence the distance function and hence KNN shows the worst performance in the case of high-dimensional data. Fritz et al. (2018) illustrated how machine learning regression trees and regression tree ensembles can be used to model interaction effects. Sonu Subudhi (2021) shows that ensemblemodels performs better than Logistic Regression, KNN, Linear Discriminant analysis, MLP Classifier for predicting the survival rate of ICU patients.

Many papers appeared in the direction of Hybrid modelling approaches to improve prediction accuracy. Bennett et al. (1998) proposed a support vector machine (SVM) approach to a

decision tree to build a Hybrid model. Kumar and Gopal's (2010) Hybrid SVM model-based decision tree and Fu Chang and Liu's (2012) decision tree as an accelerator for SVM are noticeable works in this direction. Mehmet et al. (2010) proposed a new Hybrid model based on KNN, Bayesian methods and genetic algorithm. Muhamad Safiih Lola et al. (2016) showed that the performance of multiple regression models can be improved by the Hybrid approach. They proposed an Artificial Neural Network to the MLR coins as Artificial Neural Network-Multiple Linear Regression. Gulnaz et al (2017) proposed a new classifier by incorporating the SVM learning information into the KNN and the results shows that the proposed model has good classification accuracy compared to SVM and KNN model. Tanujit Chakraborty (2019) proposed a Hybrid regression model based on a regression tree and support vector regression for boiler water quality prediction.

Presence of heteroscedasticity is more likely to occur in cross-sectional data and also in case of skewness in the distribution of regressors (Green 2000; Gujarati 2009). In the presence of heteroscedastic errors, OLS estimators are not efficient and the usual test procedures lead to an improper conclusion. It may also lead to a wider confidence interval which increases the risk of type II error (Long and Ervin 2000, Aslam 2014). There are many statistical test procedures to detect heteroscedasticity (Glejser 1969; Goldfeld and Quandt 1973; Breusch and Pagan 1979; White 1980). A novel estimation procedure proposed by Akbilgic et al (2009) is known as the Least Square Ratio Estimator (LSRE) which minimizes the sum of squares due to error in the ratio form. The LSRE shows better estimation and predictive power compared to OLSE. The OLSE can be problematic especially for outliers or non-constant variance of the error term in the model. The study of Akbilgic et al (2009) and Oznur et al (2017) shows that LSRE is robust to the presence of outliers in the model compared to OLSE.

Many studies have been carried out to remove the inter-determinate range so that the power of the DW test can be improved. Durbin and Watson (1950), Hannan (1957), Theil and Nagar

(1961), Hannan and Terrell (1968), Durbin and Watson (1971) are the few important papers based on the approximated distribution approach. Abrahamse and Koerts (1969), Farebrother (1980, 1984a, 1984b), Palm and Sneek (1989), Shively et al. (1990), Ansley et al.(1992) and Jeong and Chung (2001) works are based on numerical computation approach. Among these approaches, Jinok and Chung's Bootstrap test and Durbin and Watson $(a + bdU)$ approximation were found to be satisfactory. The power improvement by the approximated distribution approach is relatively small compared to the numerical computation approach.

The Empirical Likelihood (EL) method (Owen, 1988, 1990) owns its broad usage and fast research development to a number of important advantages. EL method combines the reliability of nonparametric methods with the effectiveness of the likelihood approach. The EL method has its unique attractions in parameter estimation and formulating goodness-of-fit tests. Art Owen (1991) extends this method to regression problems by considering fixed and random regressors for robust and heteroscedastic regressions. Jing Quin and Lawless (1994) extend the ELR confidence region when the constraints are in the form of estimating equations. Chen's (1994) EL estimators of slope and mean parameters for a simple linear regression, Jian and Shing Lau's (2000) application of the empirical likelihood method to the partially linear model, Chen and Cui's (2003) empirical likelihood for generalized linear models, Ingrid Van Keilegom's (2008) general test procedure based on EL test to specific models like Parametric models, generalized linear models are the noticeable work in the direction of application of EL to the field of Econometrics.

OBJECTIVES OF THE PRESENT STUDY

In most circumstances, if some of the independent variables are not measured directly (latent independent variables), then such variables are ignored while developing the regression model and its effect is reflected in the error term. If it is uncorrelated with independent variables, one gets an unbiased and consistent estimator from OLS estimation, but if they are correlated (which is likely) then OLS estimators are biased and inconsistent. Therefore, there is a need to include these latent variables in the regression model for improved estimators.

The Complex relationship between the variables and handling of imbalanced data in health science possesses great challenges for researchers. Most of the classification models consider only the additive nature and fail to explain the indirect effect among the variables. The inclusion of indirect effects in the regression model will improve the model's performance. Machine learning methods are extensively used to identify hidden patterns through a predictive model.

The K- nearest Neighbours model (KNN) and Regression tree (RT) are the simple, efficient and commonly used models in the machine learning literature. RT can be used to model the interaction effects. The major disadvantage of the Regression tree is that it assigns the same predicted value for all the observations in a branch, whereas KNN is sensitive to irrelevant or redundant features. Since these models are most commonly used in the literature, there is a need to develop algorithms to overcome the disadvantages of KNN and RT in order to improve prediction accuracy. The main objectives of the study are

i. To develop an appropriate model to relate the dependent variables, observed independent variables and latent independent variables and discuss the estimation of model parameters.
ii. Most of the regression and classification models fail to explain the indirect effects among the variables. It is intended to study the effectiveness of the structural equation model to incorporate the indirect effect of factors on the response.

iii. To propose a Hybrid model to overcome the major disadvantages of the regression tree and k nearest neighbors regression. Also, studied the impact of the choice of distance measures choice on the performance of the Hybrid model.

iv. To propose a Hybrid regression model by incorporating the interaction effects and to improve the prediction accuracy. This approach is extended to the classification problem.

v. For an heteroscedastic regression model, a Generalized Least Square Ratio Estimator is proposed and its properties are examined.

vi. The Empirical Likelihood Ratio Test (ELRT) is proposed for autocorrelation in the Linear Regression model and examined its finite sample performance.

SUMMARY OF THE WORK DONE

The thesis is organized into eight chapters. **Chapter 1** presents a brief introduction to structural equation modeling, special cases, basic terminology, the literature review related to the research work and the objectives of the present study.

Chapter 2 deals with Linear Structural Equations model with Latent independent variables. The relationship between dependent variables, observed independent variables (some of the dependent variables may act as independent for other dependent variables) and latent independent variables is discussed. The parameters of the proposed model are estimated through a Two-Stage Least Squares (2SLS) approach. The consistency and asymptotic normality of the estimator is derived. A simulation study is carried out to examine the small sample property of the proposed 2SLS estimator for various sample sizes. The working of the proposed method is illustrated through a real-life application with diabetic patient's data.

The work presented in this chapter received Dr. M.N Gopalan Prize for Best Research Paper award (2019) by the Indian Society for Probability and Statistics (ISPS) as part of the XXXVII Annual convention.

In chapter 3, the effectiveness of Structural equation modelling (SEM) to identify the mediating variables (indirect effect) and handle the complex relationship between the variables is illustrated through a real-life application. The results show that SEM outperforms Decision Tree, Bayesian classifier, logistic classifier and KNN classifier in terms of classification accuracy. The survival probability estimated from the structural equation modelling can be used for decision-making.

The work presented in this chapter received the Best Paper award under Statistics and Data Science track (2019) at the International conference n Advanced IT, Engineering and Management, Beeri during 2019.

Chapter 4 is focused on overcoming the major disadvantages of the K-Nearest Neighbours (KNN), Regression tree (RT) model and improving the prediction accuracy. A Hybrid regression model is proposed based on RT and KNN. The performance of the proposed model is compared with KNN for 10 types of distance measures. The performance of the proposed model is also compared with RT, KNN, and Support Vector Regression (SVR) through a Monte Carlo simulation study. The simulation result indicates that the Hybrid model outperforms all other regression models and robust to the distribution assumption. The working of the proposed model is illustrated through a real-life application on air pollution data of Delhi.

This chapter also discusses the classification performance of the proposed model. The working of the proposed DT-KNN model is illustrated for four different real-life data sets; consisting of two balanced and two imbalanced cases. The result shows that for both balance and imbalanced data sets under consideration proposed model shows better classification accuracy than all other methods.

Chapter 5 deals with a Hybrid regression model based on the RT and Multiple Linear Regression (MLR) model. The performance of the hybrid model is compared with RT, KNN, MLR and Support Vector Regression (SVR) through a Monte Carlo simulation study. The simulation result indicates that the Hybrid model outperforms all other regression models irrespective of sample size. As an application, the proposed hybrid model is used to decide the most appropriate prices for cashew nuts. The results from the Hybrid regression model can be used as a guide by the farmers for fetching better prices in the market and buyers for getting a lot of ascertained quality.

This chapter also discusses the problem of improving classification accuracy with a Hybrid approach. A hybrid classification model is proposed based on the Decision tree (DT) and Logistic regression (LR) model. The simulation and empirical results show that the proposed model has better prediction accuracy than the Synthetic minority Over Sampling Technique (SMOTE) and Random Over Sampling Examples (ROSE) sampling techniques. Hence the hybrid model is an effective way to improve classification accuracy and to deal with an imbalanced data classification problem.

In Chapter 6, the estimation of the linear regression model in the case of heteroscedastic errors is considered. In this chapter, the Generalized Least Square Ratio Estimator (GLSRE) is proposed and showed that GLSRE is reduced to the Least Square Ratio Estimator (LSRE) even in the case of heteroscedasticity. A simulation study is carried out to compare the performance of the GLSRE (LSRE) with the OLS estimator and the GLS estimator under heteroscedasticity. The simulation results show that LSRE out forms the OLSE and GLSE in case of moderate to severe heteroscedasticity irrespective of sample size and in case of weak to mild heteroscedasticity performs better for relatively small samples only. GLSE performs better than OLSE and LSRE irrespective of sample size as well as the level of heteroscedasticity in case

of the small value of error variance and in case of weak to mild heteroscedasticity for large samples. Performances of these methods are also compared using a real-life application.

Chapter 7 deals with the detection of the presence of autocorrelation in regression. A distribution-free test namely the Empirical Likelihood Ratio Test (ELRT) is proposed for autocorrelation. A Monte Carlo simulation study is carried out to compare the performance of the Empirical Likelihood Ratio Test with the D-W test, $(a + bdU)$ approximation, and bootstrapped D-W test based on estimated type-I error rate and power. The results shows that the proposed ELRT considerably outperforms the original D-W test and $(a + bdU)$ approximation irrespective of sample size, level of autocorrelation and the number of regressors. The proposed test has more accurate small sample properties than the bootstrapped D-W test when (n- k) is small. The working of the proposed test is illustrated for a real-life application using Air quality data of Delhi.

The conclusion and scope for future research are given in the **last chapter**. The R codes developed for estimation and prediction of the proposed Hybrid models are presented in the appendix.

FUTURE WORK

i. In the present thesis, a hybrid model based on the Regression tree is considered. Thus, hybrid models with SEM tree may be developed to improve the prediction accuracy.
ii. An empirical likelihood ratio test for heteroscedasticity may be developed.
iii. Least square ratio ridge estimator may be developed to deal with multicollinearity.

LIST OF PUBLISHED PAPERS

i. Satyanarayana, Ismail B, Gayathri Upadhya (2018): Predicting the Survival of ICU Patients Through Structural Equation Model, *International Journal of Scientific Research in Computer Science Applications and Management Studies, Vol. 7(6),pp. 1-5* ***(UGC list)***

ii. Satyanarayana and Ismail B (2021): Hybrid Model for Improving Classification Accuracy, *Stochastic Modeling and Applications Vol. 25 (1).pp.153-159* ***(UGC care list)***

iii. Satyanarayana and Ismail B (2021): Linear structural equations with latent variables, *International journal of Agricultural and Statistical Sciences Vol.17(1), pp.293-300.* ***(Scopus indexed)***

iv. Satyanarayana and Ismail B (2023)*: A hybrid Regression model for cashew nuts price prediction, Statistics and Applications****(accepted paper, Web of science)***

v. Satyanarayana and Ismail B (2023)*:* Generalised Least Square Ratio Estimator in Heteroscedastic regression model, *Advances and Applications in Statistics*. ***(accepted paper, Web of science)***

LIST OF PRESENTED PAPERS

i. Presented a paper on **"Predicting the Survival of ICU Patients Through Structural Equation Model"** at *International conference on Advanced IT, Engineering and Management*, organized by School of Information Technology and Bio-informics and Computer Society of India, AIMIT, Mangalore during 16 - 17th November 2018.

ii. Presented a paper titled **"Hybrid Model for improving Classification Accuracy"** at *International conference on Prof. C. R Rao's School of Thought on Statistical Sciences* Organized by Department of Statistics, Pondicherry University during 28th -29th November 2020.

iii. Presented a paper titled **"Linear structural equations with latent variables",** XXXVII Annual convention of **Indian Society for Probability and Statistics (ISPS)** held at Department of Statistics, Maharshi Dayanand University, Rohtak, Haryana, during 27-30th, December 2018.

iv. Presented a paper titled **"A hybrid regressions model for cashew nuts price prediction",** at *National conference on Trends and Perspective in Statistics and Mathematics: A Multi-Disciplinary Approach* organized by St. Xavier College and Thakur College of science and Commerce, Mumbai, 25th September 2021.

v. Presented a paper titled **"Empirical likelihood ratio test for autocorrelation in regression model"** at *International conference on Emerging Trends in Statistics and Data Science* in 40th Annual convention of **Indian Society for Probability and Statistics (ISPS)** at M. D university, Rohtak during 07-10, September 2021.

vi. Presented a paper titled **"Efficient estimation of linear regression model in case of heteroscedastic errors"** at *International conference on Emerging Trends in Statistics and Data Science* in 40th Annual convention of **Indian Society for Probability and Statistics (ISPS)** at M. D university ,Rohtak during 07-10, September 2021.

AWARDS

i. Received **Best Paper Award** under statistics and data science track at "**International conference on Advanced IT, Engineering and Management"** organized IT department of AIMIT and School of Information Technology in association with Computer Society of India, Indian society for Technical education, Iconic IT Professional, during 16 - 17th November 2018 at AIMIT campus Beeri, Mangalore.

ii. Received **Dr. M.N Gopalan Prize for** B**est Research Paper** conducted as part of the XXXVII Annual convention of **Indian Society for Probability and Statistics (ISPS)** held at Department of Statistics, Maharshi Dayanand University, Rohtak, Haryana, during 27-30th, December 2018

LIST OF COMMUNICATED PAPERS

i. Satyanarayana and Ismail B (2022): An Effective Approach for Improving Classification Accuracy, *The Advances and Applications in Mathematical Sciences*

ii. Satyanarayana and Ismail B (2022): A hybrid model Regression model for improving prediction accuracy, *Electronic Journal of Applied Statistical Analysis.*

SELECTED REFERENCES:

1. Abrahamse, A.P.J., Koerts, J. (1969). A comparison between the power of the Durbin–Watson test and the power of the BLUS test. *J. Amer. Statist. Assoc.*, 64, 938–948.
2. Akbilgic, O., and E. D. Akinci. (2009). A novel regression approach: Least squares ratio. *Communications in Statistics - Theory and Methods*, 38(9):1539–45.
3. Anderson, T. W. and H. Rubin . (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, 20(1), 46-63. https://doi.org/10.1214/aoms/ 1177730090
4. Anderson, T. W. and H. Rubin. (1950). The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, 21(4), 570-582.
5. Anderson, T.W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics*, 127(1), 1-16
6. Ansley, C.F., Kohn, R., Shively, T.S. (1992). Computing p-values for the generalized Durbin–Watson and other invariant test statistics. *J. Econom*. 54, 277–300.
7. Aslam, M. (2014). Using heteroscedasticity-consistent standard errors for the linear regression model with correlated regressors. *Communications in Statistics - Simulation and Computation*, 43 (10):2353–73. doi:10.1080/03610918. 2012.750354.
8. Bennett, K.P., Blue, J. (1998). A support vector machine approach to decision trees. *IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227),* vol. 3, pp.2396–2401.
9. Breusch, T. S., and A. R. Pagan. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47 (5):1287–94. doi:10.2307/1911963.
10. Tanujit Chakraborty & Ashis Kumar Chakraborty & Zubia Mansoor. (2019). A hybrid regression model for water quality prediction .*OPSEARCH,* volume 56, pages1167–1178.
11. Chang, F., Liu, C.C (2012). Decision tree as an accelerator for support vector machines. *In: Ding, X. (ed.) Advances in Character Recognition, (Chapter 8)*, doi:10.5772/52227.
12. Chen, S.X and Cui, H. (2003).An extended empirical likelihood for generalized linear models. *Statistical Sinica*, Vol. 13, No. 1 , pp. 69-81.
13. Chen. (1994). Empirical likelihood confidence interval for linear regression coefficient. *Journal of Multivariate analysis*, 49.29-40.
14. Durbin, J., Watson, G.S. (1950). Testing for serial correlation in least square regression I. *Biometrika*, 37, 409 – 428.

15. Durbin, J., Watson, G.S. (1971). Testing for serial correlation in least square iii. *Biometrika*, 58(1), 1–19.
16. Farebrother, R.W. (1980). Algorithm AS153. Pan's procedure for the tail probabilities of the Durbin– Watson statistic. *Appl. Statist*., 29, 224 – 227.
17. Farebrother, R.W. (1984a). The distribution of a linear combination of central 12 random variables. A remark on AS153: Pan's procedure for the tail probabilities of the Durbin–Watson statistic. *Appl. Statist*.,33, 363–366.
18. Farebrother, R.W. (1984b). A remark on Algorithm AS106, AS153 and AS155: the distribution of a linear combination of 12 random variables. *Appl. Statist.*,33, 366 –369.
19. Fritz Schiltz, Chiara Masci, Tommaso Agasisti & Daniel Horn. (2018). Using regression tree ensembles to model interaction effects: a graphical approach. *Applied Economics*, 50:58, 6341-6354, DOI: 10.1080/00036846.2018.1489520.
20. Glejser, H. (1969). A new test for heteroskedasticity. *Journal of the American Statistical Association*, 64 (325):316–23. doi:10.1080/01621459.1969.10500976.
21. Goldfeld, S. M., and R. E. Quandt.(1973). A Markov model for switching regressions. *Journal of Econometrics*, 1 (1): 3–15. doi:10.1016/0304-4076(73)90002-X.
22. Green, W. (2000). Econometric analysis. *New York: Pearson*
23. Gujarati, D. N. (2009). Basic econometrics. *New Delhi, India: Tata McGraw-Hill Education*
24. Gulnaz Alimjan., Tieli Sun., Hurxida Jumahun. and Yu Guan., Wanting Zhou. and Hong-guang Sun.(2017). A Hybrid Classification Approach Based on Support Vector Machine and K-Nearest Neighbor for Remote Sensing Data. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(10) ,1750034.
25. Hannan, E.J., (1957). Testing for serial correlation in least squares regression. *Biometrika*, 44, 57– 66.
26. Hannan, E.J., Terrell, R.D.(1968). Testing for serial correlation after least squares regression. *Econometrica*, 36, 133–150.
27. Ibomoiye Domor Mienyea, Yanxia Suna and Zenghui Wangb.(2019). Prediction performance of improved decision tree-based algorithms: a review. *2nd International Conference on Sustainable Materials Processing and Manufacturing (SMPM 2019)*.
28. Ingrid Van Keilegram. (2008). Empirical Likelihood-based regression testing. *Electronic journal of Statistics*,Vol.2, 581-604.
29. Jian Shi Tai-Shing Lau. (2000). Empirical Likelihood for Partially Linear Models. *Journal of Multivariate analysis*. Vol.72 (1), 132-148

30. Jing Quin, Jerry Lawless. (1994). Empirical Likelihood and General Estimating Equations. *Ann. Statist*. 22 (1), 300 – 325
31. Jinook Jeonga, Seoung Chung. (2001). Bootstrap tests for autocorrelation. *Computational Statistics & Data Analysis* ,38 ,49–69
32. Joreskog, K.G. (1970). A general method for analysis of covariance structures. *Biometrika*, 57, 239-251.
33. Kumar, M.A., Gopal, M. (2010). A hybrid SVM based decision tree. *Pattern Recogn*., 43(12), 3977–3987.
34. Long, J. S., and L. H. Ervin. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54 (3):217–24.
35. Mehmet Aci., Cigdem Inan., and Mutlu Avci. (2010).A hybrid classification method of k-nearest neighbor, Bayesian methods and genetic algorithm. *Expert Systems with Applications*, 37,5061–5067.
36. Muhamad Safiih, Mohd Noor Afiq Ramlee, G. Sugan Gunalan, Nurul Hila Zainuddin, Razak Zakariya, Md Suffian Idris, Idham Khalil. (2016). Improved the Prediction of Multiple Linear Regression Model Performance Using the Hybrid Approach: A Case Study of Chlorophyll-a at the Offshore Kuala Terengganu. *Open Journal of Statistics*, 2016, 6, 789-804.
37. Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18, 90–120.
38. Owen, A.(1988): Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2), 237-249.
39. Owen, A.(1991): Empirical likelihood for linear models. *Annals of Statistics*,19,1725-1747.
40. Oznur Işçi Guneri, Atilla Goktas. (2017). A Comparison of Ordinary Least Squares Regression and Least Squares Ratio via Generated Data. *American Journal of Mathematics and Statistics*, 7(2): 60-70 DOI: 10.5923/j.ajms.20170702.02
41. Palm, F.C., Sneek, J.M. (1984). Significance tests and spurious correlation in regression models with autocorrelated errors. *Statist. Hefte*, 25, 87–105.
42. Sargan, J.D. (1958). Estimation of economic relationships using instrumental variables. *Econometrica*, 67, 557- 586.
43. Shively, T.S., Ansley, C.F., Kohn, R. (1990) . Fast evaluation of the distribution of the Durbin–Watson and other invariant test statistics in time series regression. *J. Amer. Statist. Assoc*., 85, 676 – 685.

44. Subudhi, S., Verma, A., Patel, A.B. *et al.* Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *npj Digit. Med.* **4**, 87 (2021). https://doi.org/10.1038/s41746-021-00456-x
45. T. Mitchell .(1997). Machine learning, *McGraw-Hill, New York.*
46. Theil, H. (1953a). Repeated least-squares applied to complete equation systems. *Central Planbureau Memorandum.*
47. Theil, H. (1953b). Estimation and simultaneous correlation in complete equation systems. Central Planbureau Memorandum. *(Reprinted In: Raj, B., Koerts, J. (Eds.), Henri Theil's Contributions to Economics and Econometrics, Vol. 1, Kluwer, Dordrecht.)*
48. Theil, H., 1954. Estimation of parameters in econometric models. *Bulletin of the International Statistical Institute,*34, 122–129.
49. Theil, H., Nagar, A.L. (1961). Testing the independence of regression disturbances. *J. Amer. Statist. Assoc.*, 56, 793–806.
50. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48 (4):817–38.

Satyanarayana
(Research Scholar)

Prof. Ismail B.
(Research Guide)